Learning Spoken Document Similarity and Recommendation using Supervised Probabilistic Latent Semantic Analysis

K. Thambiratnam, F. Seide

Microsoft Research Asia, 5F Beijing Sigma Center,
No. 49 Zhi Chun Rd., 100080 Beijing, P.R.C.
[kit,fseide]@microsoft.com

Abstract

This paper presents a model-based approach to spoken document similarity called Supervised Probabilistic Latent Semantic Analysis (PLSA). The method differs from traditional spoken document similarity techniques in that it allows similarity to be *learned* rather than approximated. The ability to learn similarity is desirable in applications such as Internet video recommendation, in which complex relationships like user-preference or speaking style need to be predicted. The proposed method exploits prior knowledge of document relationships to learn similarity. Experiments on broadcast news and Internet video corpora yielded 16.2% and 9.7% absolute mAP gains over traditional PLSA. Additionally, a cascaded Supervised+Discriminative PLSA system achieved a 3.0% absolute mAP gain over a Discriminative PLSA system, demonstrating the complementary nature of Supervised and Discriminative PLSA training.

Index Terms: Document Similarity, Document Recommendation, Probabilistic Latent Semantic Analysis, Spoken Document Retrieval, Information Retrieval

1. Introduction

Consuming multimedia content on the Internet is hugely popular through media such as hosted user-generated content, premium streaming video, and podcasts. However, discovering such content continues to be a challenge. Currently, video portals and search engines are the primary modes of access, but these suffer from user experience issues. Video portals require highly interactive browsing that can detract from the overall experience of viewing desired content. Search engines require users to have a clear intention of the content they wish to consume. Instead, what is needed are techniques for enabling content discovery that limit user interaction while providing relevant content.

One such solution is a recommendation system. Recommendation systems suggest content to a user using methods such as content similarity, user profiling [1, 2], and collaborative filtering [3, 4]. To date, collaborative filtering is the most popular method for recommendation, but requires a large amount of user history data. For spoken content, spoken document similarity is an alternative means of recommendation that does not require user data. Spoken document similarity can also be used for organizing and linking documents in a large multimedia database to facilitate structured browsing.

The goal of spoken document similarity is to use the entire spoken content of a multimedia document as a query for retrieving similar or relevant documents from a database. Typically Automatic Speech Recognition (ASR) is first used to generate transcriptions. Text similarity techniques such as Vector Space Modeling (VSM) [5], Latent Semantic Analysis (LSA) [6] or Probabilistic Latent Semantic Analysis (PLSA) [7], can then be used to determine similarity. Unfortunately, high ASR word error rates degrade spoken document similarity performance which means further work remains to be done on *spoken* document similarity.

A limitation of traditional spoken document similarity techniques is that the *type* of similarity is typically skewed towards topical similarity. Unfortunately, spoken similarity can be defined in other ways, such as speaking style or language choice (eg. contemporary versus traditional language). The ability to predict a variety of similarity types is desirable, particularly for entertainment domains such as Internet video recommendation.

Discriminative PLSA was proposed in [8] to enable the prediction of different similarity types. This method used discriminative training to train a similarity model to mimic the relationships encoded within a document relationship matrix. Notable gains in performance were achieved over traditional similarity approaches. However, this technique only trained document-specific parameters. Thus global parameters, such as the word-to-factor PDF $p(w_k|z_j)$ in PLSA, remained unchanged. As a result, if the seed model for discriminative training had poorly estimated global parameters, the resulting discriminatively trained model would continue to use these poor estimates.

This work presents a method for exploiting relationship information *directly* within PLSA training. It allows updating of both document-specific and global parameters. The maximum likelihood PLSA training criterion is augmented with the document relationship matrix. Model training thus becomes supervised, and tries to maximize similarity prediction given the relationships within the training data. A side-effect of this is that the training objective function is better aligned with the evaluation objective function, resulting in a more robust model.

The paper is organized as follows. Background theory is briefly discussed in Section 2. This is followed by details of the proposed Supervised PLSA training algorithm in Section 3. Experiments and results are reported in Section 4 and conclusions and future work are detailed in Section 5.

2. Background

Arguably one of the most common similarity techniques is VSM [5], or commonly known as TFIDF. In VSM, each document is represented by a document vector $\mathbf{x^i}$, where $x_k^i = TF(d_i, w_k) \times \sqrt{IDF(w_k)}$ represents the relative frequency of word w_k in document d_i. Here $TF(d_i, w_k) = p(w_k|d_i) = n(d_i, w_k)/\sum_{k'=1}^{K} n(d_i, w_{k'})$ is the intra-document word frequency or Term Frequency, $IDF(w_k) = \log D/N_D(w_k)$ is

the well known Inverse Document Frequency (IDF) global term weighting, $n(d_i, w_k)$ is the number of occurrences of w_k in document d_i, D is the number of database documents, and $N_D(w_k)$ is the number of documents in the database in which word w_k occurs at least once. Document similarity is then evaluated using the cosine distance measure, SIM_{VSM}:

$$SIM_{VSM}(\mathbf{x^1}, \mathbf{x^2}) = \frac{\mathbf{x^1} \cdot \mathbf{x^2}}{|\mathbf{x^1}|\,|\mathbf{x^2}|} \quad (1)$$

2.1. Probabilistic Latent Semantic Analysis

Probabilistic Latent Semantic Analysis, proposed by [7], is a probabilistic approach to Latent Semantic Analysis [6]. Both methods represent documents within a semantic factor space, allowing co-occurrence, polysemy and synonymy to be exploited for document similarity. PLSA is a probabilistic approach and thus is more attractive for the speech domain.

PLSA attempts to capture word co-occurrence patterns by modeling the document-word co-occurrence matrix using the joint distribution, $p(d, w)$, and the latent semantic variable, z:

$$p(d_i, w_k) = \sum_{j=1}^{J} p(z_j)p(d_i|z_j)p(w|z_j) \quad (2)$$

The individual PDFs, $p(z_j)$, $p(d_i|z_j)$ and $p(w_k|z_j)$ are trained using Expectation Maximization (EM). This enables a number of interesting semantic representations, including the factor-space representation of each document $\mathbf{d^i}$, where $d^i_j = p(z_j|d_i)$ is the Expected Factor Frequency (EFF) (*expected* here emphasizes the fact that $\mathbf{d^i}$ is only a probabilistic estimate of the true factor frequency vector). Note the similarity here with the TF document vector, since TF can be written as $p(w_k|d_i)$. Document similarity can thus be computed using VSM [7] by computing the cosine distance of the EFF vectors:

$$SIM_{PVSM}(\mathbf{d^1}, \mathbf{d^2}) = \frac{\mathbf{d^1} \cdot \mathbf{d^2}}{|\mathbf{d^1}|\,|\mathbf{d^2}|} \quad (3)$$

The above similarity measure requires an EFF representation for unseen query documents in order to compute similarity with documents within a database. Typically, a query document, q, is approximated or *folded* into the factor space using the PLSA PDFs. One approach is to fix $p(w_k|z_j)$ and $p(z_j)$ and to then use EM to estimate $p(z_j|q)$ [7]. Here, the empirical distribution $\tilde{p}(q, w_k)$ is approximated using the query word counts $n(q, w_k)$. Alternately, the empirical word distribution $\tilde{p}(w_k|q)$ derived from the query TF vector $\mathbf{y}$, $y_k = p(w_k|q) = TF(q, w_k)$ can be used to probabilistically predict a query's representation, using $p(z_j|q) = \sum_{k=1}^{K} p(z_j|w_k)\tilde{p}(w_k|q)$ assuming that $p(z_j|w_k, q) \approx p(z_j|w_k)$.

2.2. Discriminative PLSA

A limitation of both VSM and PLSA is that prior knowledge of document relationships is not exploited. Thus, similarity is computed completely blindly without any *learned* understanding of the similarity that is trying to be predicted. Discriminative PLSA was proposed in [8] and allowed similarity to be *learned* rather than *guessed* by using examples of known relationships and non-relationships. A discriminative training framework was used to train an ensemble of document similarity models that mimicked the relationships encoded in a document relationship matrix. Additionally, a Minimum Classification Error (MCE) training criterion was used to enable better learning. A brief introduction is given below.

A document is modeled using $\lambda^{\mathbf{i}}$ where $\lambda^i_j = d^i_j \gamma^i_j$ and d^i_j is the EFF. γ^i_j is an importance weight that reflects the importance of a factor j for discrimination of document i. The factor importance weights are similar to global term weights, such as IDF weights, but are trained on a per-document basis. This results in the importance weighted VSM similarity measure

$$SIM_{GVSM}(\mathbf{d^1}, \mathbf{d^2}) = \frac{(\gamma^{\mathbf{1}} \otimes \mathbf{d^1}) \cdot \mathbf{d^2}}{|\gamma^{\mathbf{1}} \otimes \mathbf{d^1}|\,|\mathbf{d^2}|} \quad (4)$$

where $\otimes$ is element-by-element vector multiplication. Model training is then performed as follows. First a training set is constructed using a set of target document word vectors, $\mathbf{R} = (\mathbf{r_1}, \ldots, \mathbf{r_D})$, a set of training query document vectors, $\mathbf{X} = \{\mathbf{x_1}, \ldots, \mathbf{x_N}\}$, and the $D \times N$ similarity matrix, Ψ, where Ψ_{ij} is a binary function indicating whether target document i is similar to training query j. An initial set of target models, $\mathbf{\Lambda}$, is bootstrapped appropriately (eg. using a constant for all γ^i_j and training other parameters using PLSA). Discriminative training is then performed iteratively using the update equations in [8].

3. Supervised PLSA

A limitation of Discriminative PLSA is its reliance on a seed model. Since only the document-specific parameters, $\gamma^{\mathbf{i}}$, are trained, poor seed model estimates of the global parameters $p(w_k|z_j)$ and $p(z_j)$ will continue to be poor after discriminative training. Unfortunately, it is not possible to discriminatively train the PLSA global parameters without introducing strong constraints that limit performance gains.

Thus, this work presents a technique for exploiting prior document relationship information *directly* within PLSA training. Supervised Probabilistic Latent Semantic Analysis allows maximum likelihood training of document-specific and global parameters with respect to a similarity objective function. The primary advantage of this technique over traditional PLSA is that training is done in a supervised fashion. Thus, task specific similarity prediction can actually be *learned* from the data. Additionally, unlike traditional PLSA, the training criterion in supervised PLSA is aligned with the evaluation criterion.

Supervised PLSA training is performed as follows. A training set is constructed using a set of target document vectors, $\mathbf{X} = \left(\mathbf{x^1}, \ldots, \mathbf{x^D}\right)$, a set of training query vectors, $\mathbf{R} = \left\{\mathbf{r^1}, \ldots, \mathbf{r^N}\right\}$, and the $D \times N$ similarity matrix, Ψ, where Ψ_{iq} is a binary function indicating if document $\mathbf{x^i}$ is similar to query $\mathbf{r^q}$. The probability of a target document, $\mathbf{x^i}$ being related to query $\mathbf{r^q}$ can then be expressed as the joint probability

$$p(\mathbf{x^i}, \mathbf{r^q}|\Lambda) \quad = \quad \prod_k p(\mathbf{x^i}, w_k, \mathbf{r^q}|\Lambda)^{n(\mathbf{x^i}, w_k)} \quad (5)$$

where $n(\mathbf{x^i}, w_k)$ is the count of w_k in $\mathbf{x^i}$. This probability can be expressed in terms of the individual z_j factor contributions

$$p(\mathbf{x^i}, w_k, \mathbf{r^q}) \quad = \quad \sum_j p(z_j, \mathbf{x^i}, w_k, \mathbf{r^q}|\Lambda) \quad (6)$$

$$p(z_j, \mathbf{x^i}, w_k, \mathbf{r^q}|\Lambda) = p(d_i|z_j)p(z_j|w_k)p(w_k|\mathbf{r^q})p(\mathbf{r^q}) \quad (7)$$

given the appropriate independence assumptions. Then the total log likelihood of target set X, training query set R, and the similarity matrix Ψ, is

$$\log p(\mathbf{X}, \mathbf{\Psi}, \mathbf{R}|\Lambda) = \log \prod_{i=1}^{N} \prod_{q}^{D} \Psi_{iq} \times p(\mathbf{x^i}, \mathbf{r^q}|\Lambda) \quad (8)$$

This log likelihood can be maximized using the EM algorithm as follows. The training parameter set is introduced: $s_i^j = p(d_i|z_j)$, $t_k^j = p(w_k|z_j)$, $u_j = p(z_j)$ and $r_k^q = p(w_k|\mathbf{r_q})$. Additionally, a set of constraints are introduced to enforce a PDF.

$$\forall j : \sum_i s_i^j = 1, \quad \forall j : \sum_k t_k^j = 1, \quad \sum_j u_j = 1 \tag{9}$$

Then EM theory states that the total likelihood in Eqn. 8 can be maximized under the constraints in equation set 9 by maximizing the auxiliary function

$$\begin{aligned} Q(\Lambda) = E_z \left[\log p\left(\mathbf{X}, \mathbf{\Psi}, \mathbf{R}|\mathbf{\Lambda}\right) \middle| \mathbf{X}, \Psi, \mathbf{R} \right] \\ -\gamma \left[\sum_j p(z_j) - 1 \right] - \sum_j \alpha_j \left[\sum_i p(d_i|z_j) - 1 \right] \\ - \sum_j \beta_j \left[\sum_i p(w_k|z_j) - 1 \right] \end{aligned} \tag{10}$$

where the constraints have been including using the Lagrange multipliers α_j, β_j, and γ. Solving this gives the E-step update

$$p(z_j|\mathbf{x^i}, \mathbf{r^q}, \Lambda) = \frac{\sum_k n(\mathbf{x^i}, w_k) p(z_j, \mathbf{x^i}, w_k, \mathbf{r^q}|\Lambda)}{\sum_j \sum_k n(\mathbf{x^i}, w_k) p(z_j, \mathbf{x^i}, w_k, \mathbf{r^q}|\Lambda)} \tag{11}$$

and M-step parameter updates

$$s_i^{j*} = \frac{\sum_q \sum_k n(\mathbf{x^i}, w_k) p(z_j|\mathbf{x^i}, \mathbf{r^q}) \Psi_{iq}}{\sum_{i'} \sum_q \sum_k n(\mathbf{x^{i'}}, w_k) p(z_j|\mathbf{x^{i'}}, \mathbf{r^q}) \Psi_{i'q}} \tag{12}$$

$$t_k^{j*} = \frac{\sum_q \sum_i n(\mathbf{x^i}, w_k) p(z_j|\mathbf{x^i}, \mathbf{r^q}) \Psi_{iq}}{\sum_i \sum_q \sum_{k'} n(\mathbf{x^i}, w_{k'}) p(z_j|\mathbf{x^i}, \mathbf{r^q}) \Psi_{iq}} \tag{13}$$

$$u_j{}^* = \frac{\sum_i \sum_q \sum_k n(\mathbf{x^i}, w_k) p(z_j|\mathbf{x^i}, \mathbf{r^q}) \Psi_{iq}}{\sum_{j'} \sum_i \sum_q \sum_k n(\mathbf{x^i}, w_k) p(z_{j'}|\mathbf{x^i}, \mathbf{r^q}) \Psi_{iq}} \tag{14}$$

4. Experiments and Results

Experiments were performed to evaluate similarity prediction for Supervised PLSA. Evaluations were conducted on two data sets: 1) TDT2 - a broadcast news corpus with topically homogeneous segments, and 2) NetVid - a collection of various styles of Internet video clips, including news, sports, documentary and entertainment clips.

4.1. Experiment setup

The TDT2 data was taken from the well-known TDT2 corpus. A 68 hour subset of speech was selected as the target document set. ASR transcripts were then used to create document vectors for each target document. The similarity matrix for Supervised PLSA and Discriminative PLSA was built using topic annotations from 91 topic groups in the TDT2 corpus. Documents were marked as related if they shared at least one topic label. An evaluation query set was also constructed using 41 additional query documents from TDT2 that shared a topic label with at least one document in the target document set. On average, each query had 15 related documents in the target database.

The NetVid set was constructed using a diverse collection of Internet video clips, including news, sports, documentary and entertainment clips. A total of 247 hours of videos (4800 clips) was used as the target document set. ASR transcripts were used to create document vectors for each of these documents. However, as the word error rate was $\approx 30\%$ absolute worse than the TDT2 set, document vectors were constructed using only words with high posterior scores (> 0.8), resulting in only partial document transcripts being used for document vectors. The similarity matrix was constructed using video tag information. Each video had about 3-8 manually assigned tags and documents were considered related if they shared at least one tag. An evaluation query set was also constructed using 150 additional videos that shared at least one tag with a document in the target set. On average, each query had 240 related documents.

Method	Recall @ X% Precision		mAP
	@90%	@80%	
TFIDF	22.1	35.7	53.9
PVSM	1.5	7.17	37.0
PVSM $\oplus$ GVSM	33.2	45.9	66.0
SUP-PVSM	18.3	30.4	53.2
SUP-PVSM $\oplus$ GVSM	27.9	49.0	69.0

Table 1: Similarity prediction results on TDT2 corpus. PLSA systems used 200 factors. $X \oplus$ GVSM indicates that model X was used as the seed model for discriminative training

Whereas the relationships in the TDT2 corpus were based on spoken topical similarity, those in the NetVid corpus were more abstract, including relationships such as belonging to the same episode of a series, or originating from the same source. Additionally the number of relationships per NetVid document was considerably higher since relationships could be formed across multiple tags. Thus, it was expected that similarity prediction would be considerably more challenging for the NetVid corpus.

Performance was evaluated for 3 baseline systems, 1) TFIDF: VSM using TFIDF, 2) PVSM: standard PLSA with SIM_{PVSM} similarity metric, and 3) PVSM $\oplus$ GVSM: Discriminative PLSA with SIM_{GVSM} similarity metric. All PLSA systems used EFF prediction for query folding.

Spoken document similarity was evaluated as a classification task. For each document in the evaluation query set, all documents in the target document set were scored. Precision, Recall and Mean Average Precision (mAP) were then used for comparing system performance.

4.2. Unsupervised versus Supervised PLSA

Experiments were first performed on the TDT2 corpus to compare the performance of Supervised PLSA with traditional unsupervised PLSA. The results of these experiments are shown in Table 1. It was found that the Supervised PLSA system yielded a notable 16.2% absolute mAP gain over the unsupervised PLSA system. Gains were consistent across all operating points (10% and 20% miss rate points are shown in Table 1).

The benefits of Supervised PLSA can in part be attributed to a better trained word-to-factor distribution, $p(w_k|z_j)$. In order to demonstrate this, the word factorization agreement was measured for each model. A higher agreement would indicate that words from the same semantic concept were being probabilistically mapped to a greater number of common factors.

The coefficient of agreement was defined as $\rho(\mathbf{W}, \Lambda) = \frac{1}{J}\left[\sum_j^J \prod_n^N \log p(z_j|w_n)\right]^{1/N}$, where the set of topically related words, $\mathbf{W} = (w_1, \ldots, w_N)$ was chosen empirically. The coefficients for a selection of word sets are shown in Table 2. It was found that *in general* (though not always), there was a statistically greater agreement for Supervised PLSA.

Topic	Words	PVSM	SUP-PVSM
cars	car, motor, drive speed, auto, race	18.7	19.2
weather	weather, storm, rain, cloud, front, forecast	19.1	19.5
stocks	stock, trade, share, market, business, price	18.3	18.9

Table 2: Coefficient of agreement for unsupervised PLSA (PVSM) vs. Supervised PLSA (SUP-PVSM)

4.3. Experiments on TDT2 Broadcast News

Experiments were also performed to compare Supervised PLSA to the other baseline systems. The results are also shown in Table 1. It was found that Supervised PLSA performance was slightly below TFIDF performance, and considerably poorer than PVSM $\oplus$ GVSM. It was expected though that PVSM $\oplus$ GVSM would outperform Supervised PLSA since the discriminatively trained system exploited both *related* documents and *unrelated* documents during training.

However, a notable gain in performance over PVSM $\oplus$ GVSM (the best baseline system) was observed when the Supervised PLSA model was used as the seed model for discriminative training. This system, SUP-PVSM $\oplus$ GVSM, achieved a 3.0% absolute gain in mAP and ranked as the best system for the TDT2 evaluations. The results indicate that using better initial parameter estimates for discriminative training definitely translates to improvements in final performance. Thus, Supervised PLSA and Discriminative PLSA are complementary training algorithms that can be cascaded.

4.4. Experiments on NetVid Internet Video

Evaluations were also performed on the NetVid corpus to measure performance when trying to predict more complex similarities using spoken document similarity. The results for these evaluations are shown in Table 3. Most notable is that performances for the TFIDF and PLSA systems were considerably poorer than for the TDT2 corpus, with absolute drops in mAP of 37.3% and 25.9% respectively. This drop in performance can be attributed to a combination of trying to predict more complex similarities, using more erroneous ASR transcripts, and using a much larger target database. However, the reduction in performance for PVSM $\oplus$ GVSM was much less, with only a 6.3% absolute reduction in mAP.

Supervised PLSA once again achieved a notable gain over unsupervised PLSA, with an absolute mAP gain of 9.7%. Additionally, it also outperformed the TFIDF system. Gains were also observed when using the Supervised PLSA model as a seed for discriminative training. The SUP-PVSM $\oplus$ GVSM achieved a 2.6% absolute gain in mAP over the PVSM $\oplus$ GVSM system, and a particularly pleasing 5.8% gain at 90% precision. Although it is not shown here, the gains of SUP-PVSM $\oplus$ GVSM were consistent across operating points, except for a slight degradation at 75-80% precision.

Overall, the reported experiments demonstrated that Supervised PLSA was able to achieve significant gains over unsupervised PLSA. In addition, when combined with discriminative training, a cascaded Supervised PLSA and Discriminative PLSA system was able to yield the best performance on both evaluated tasks.

Method	Recall @ X% Precision		mAP
	@90%	@80%	
TFIDF	2.0	2.5	16.6
PVSM	1.9	2.4	11.1
PVSM $\oplus$ GVSM	19.8	35.5	59.7
SUP-PVSM	3.8	5.7	20.8
SUP-PVSM $\oplus$ GVSM	25.6	34.2	62.3

Table 3: Similarity prediction results on NetVid corpus. PLSA systems used 1000 factors. $X \oplus$ GVSM indicates that model X was used as the seed model for discriminative training

5. Conclusion

This paper has presented a new model-based approach to spoken document similarity called Supervised PLSA. The proposed technique allows prior knowledge of document relationships to be incorporated into the PLSA training process and was shown to yield significant improvements for spoken document similarity prediction. Experiments on the broadcast news TDT2 corpus and the Internet video NetVid corpus yielded 16.2% and 9.7% absolute mAP gains over traditional PLSA. Additionally, it was shown that Supervised PLSA was complementary with Discriminative PLSA, with a cascaded Supervised+Discriminative PLSA trained model achieving a 3.0% gain over Discriminative PLSA. This cascaded training approach achieved the best performance on all evaluated tasks.

6. Acknowledgments

The author would like to thank Roger Yu for his invaluable contributions during research discussions.

7. References

[1] K. Lang, "Newsweeder: Learning to filter NetNews," in *Machine Learning. Proceedings. 12th International Conference on*, 1995.

[2] B. Krulwich and C. Burkey, "Learning user information interests through extraction of semantically significant phrases," in *Machine Learning in Information Access. Proceedings. AAAI Spring Symposium on*, 1996.

[3] *Visualizing the Semantic Web*, chapter Recommender Systems for the Web., Springer Verlag, 2002.

[4] B. M. Sarwar, G. Karypis, J. A. Konstan, and J. Riedl, "Item-based collaborative filtering recommendation algorithms," in *10th World Wide Web, 2001. Proceedings. International Conference on*, 2001.

[5] C. D. Manning and H. Schutze, *Foundations of Statistical Natural Language Processing*, chapter Topics in Information Retrieval, pp. 539–41, MIT Press, 1999.

[6] S. C. Deerwester, S. T. Dumais, T. K. Landauer, G. W. Furnas, and R. A. Harshman, "Indexing by Latent Semantic Analysis," *Journal of the American Society of Information Science*, vol. 41, no. 6, pp. 391–407, 1990.

[7] Thomas Hofmann, "Probabilistic Latent Semantic Analysis," in *Proc. of Uncertainty in Artificial Intelligence, UAI'99*, Stockholm, 1999.

[8] K. Thambiratnam, P. Yu, and F. Seide, "Discriminatively Trained Spoken Document Similarity Models and their Application to Probabilistic Latent Semantic Analysis," in *Spoken Language Technology, 2006. Proceedings. IEEE/ACL International Workshop on*, 2006.